


Predicting the Basic Level in a Hierarchy of Concepts

Laura Hollink ^[0000-0002-6865-0021], Aysenur Bilgin^[0000-0002-6225-9953], and
Jacco van Ossenbruggen^[0000-0002-7748-4715]

Centrum Wiskunde & Informatica, Amsterdam, The Netherlands
{laura.hollink, aysenur.bilgin, jacco.van.ossenbruggen}@cwi.nl

Abstract. The “basic level”, according to experiments in cognitive psychology, is the level of abstraction in a hierarchy of concepts at which humans perform tasks quicker and with greater accuracy than at other levels. We argue that applications that use concept hierarchies could improve their user interfaces if they ‘knew’ which concepts are the basic level concepts. This paper examines to what extent the basic level can be learned from data. We test the utility of three types of concept features, that were inspired by the basic level theory: lexical features, structural features and frequency features. We evaluate our approach on WordNet, and create a training set of manually labelled examples from different part of WordNet. Our findings include that the basic level concepts can be accurately identified within one domain. Concepts that are difficult to label for humans are also harder to classify automatically. Our experiments provide insight into how classification performance across different parts of the hierarchy could be improved, which is necessary for identification of basic level concepts on a larger scale.

1 Introduction

Applications that use metadata rely on knowledge organization systems – taxonomies, thesauri, ontologies and more recently knowledge graphs – to provide controlled vocabularies and explicit semantic relationships among concepts [27]. While these various knowledge organization systems (KOSs) may use different formal languages, they all share similar underlying data representations. They typically contain instances and classes, or concepts, and they use subsumption hierarchies to organize concepts from specific to generic.

In this paper, we aim to enrich the concept hierarchy of a widely used KOS, WordNet, by predicting which level of abstraction is the ‘basic level.’ This is a notion from the seminal paper by Rosch et al. [22] in which they present the theory of “basic level categories.”¹ The core idea is that in a hierarchy of concepts there is one level of abstraction that has special significance for humans.

¹ Note that vocabulary varies per research community and throughout time. Rosch’s “categories” would be called “classes” or “concepts” in recent Knowledge Representation literature.

At this level, humans perform tasks quicker and with greater accuracy than at superordinate or subordinate levels. In a hierarchy of edible fruits, this so called ‘basic level’ is at the level of *apple* and not at the levels of *granny smith* or *fruit*; in a hierarchy of tools it is at the level of *hammer* rather than *tool* or *maul*; and in a hierarchy of clothing it is at the level of *pants*. In a series of experiments, Rosch demonstrated that humans consistently display ‘basic level effects’ – such as quicker and more accurate responses – across a large variety of tasks.

In contrast, in current knowledge graphs (and other KOSs) each level in the hierarchy is treated equally. To illustrate why this may be problematic, consider the following example. Using a taxonomy of fruits combined with a metadata record saying that an image displays a granny smith, we can infer new facts: that it displays an apple and that it displays a fruit. However, that doesn’t tell us which is the best answer to the question “What is depicted?” – a granny smith, an apple or a fruit? In cases where the concept hierarchy is deep, there might be dozens of concepts to choose from, all equally logically correct descriptions of the image. KOSs generally have no mechanism for giving priority to one level of abstraction over another.

We argue that applications that use knowledge graphs could significantly improve their user interfaces if they were able to explicitly use basic level concepts when appropriate. In other words, if they were able to predict for which concepts in the graph users can be expected to display basic level effects. In the example above, we illustrated how computer vision systems could be designed to describe the objects they detect at the basic level rather than at subordinate or superordinate levels, so that users can react as quickly as possible. Another example is an online store, that could cluster products at the basic level to give users a quick overview of what is sold, rather than choosing more specific or more general clusters. Automatic summarization systems could describe the contents of an article at the basic level, etc. It is important to note that we do not argue that the basic level should *always* be the preferred level of abstraction. For example, in indexing as it is done in digital libraries, it is often advisable to select concepts as specific as possible. Also, in application-to-application situations where there is no interaction with a human user, basic level effects are obviously irrelevant.

Motivated by these example scenarios, our goal is to predict which concepts in a given concept hierarchy are at the basic level. We do this in a data-driven manner, in contrast to the laboratory experiments with human users that have been conducted in cognitive psychology.

We train a classifier using three types of features. Firstly, we elicit lexical features like word-length and the number of senses of a word, since it has commonly been claimed that basic level concepts are denoted by shorter and more polysemous words [18,25,8]. Secondly, we extract structural features, such as the number of subordinates of a concept and the length of its description. This is motivated by a definition of the basic level being “the level at which categories carry the most information” [22]. Finally, we obtain features related to the frequency of use of a word, since basic level concepts are thought to be used often by humans.

To test our approach, we apply it to the concept hierarchy of WordNet, a widely used lexical resource, and classify WordNet concepts as basic level or not-basic level. For training and testing, we create a gold standard of 518 manually labelled concepts spread over three different branches of the WordNet hierarchy.

Frequency features are extracted from Google Ngram data. Lexical and structural features are extracted from the concept hierarchy itself, i.e. from WordNet. In a series of experiments, we aim to answer three research questions: 1) to what extent can we predict basic level concepts within and across branches of the hierarchy

, 2) how can we predict the basic level in new, previously unseen parts of the hierarchy, and 3) how does machine classification compare to human classification, i.e. what is the meaning of disagreement between human annotators? We believe the answer to these three questions will bring us one step closer to the overall aim of being able to predict the basic level on a large scale in all kinds of concept hierarchies, helping applications built on top of them to interact with users more effectively.

All data is publicly available²: the gold standard of manually labelled concepts, the annotation protocol, code used for feature extraction, as well as an RDF dataset of all predicted basic level concepts in WordNet.

2 Background: the Basic Level

Rosch et al. [22] demonstrated basic level effects across a large variety of tasks. For example, they found that people, when asked to verify if an object belonged to a category, reacted faster when it was a basic level category (“Is this a chair” is answered quicker than “Is this furniture?”); when asked to name a pictured object, people chose names of basic level concepts (They said “It is an apple” rather than “It is a golden delicious”); and when asked to write down properties of a concept, people came up with longer lists if the concept was at the basic level (many additional properties were named for “car” compared to the properties of its superordinate “vehicle”, while few additional properties were mentioned for “sports car”). In the present paper, we aim to derive ‘basic levelness’ in a data driven manner, rather than by performing psychological experiments, to allow for enrichment of concept hierarchies at a large scale.

Rosch’s initial experiments were done on a relatively small set of nine hierarchies of ten concepts each. She chose common, tangible concepts, such as fruits, furniture and musical instruments. Later, basic level effects were also demonstrated in other types of concepts, such as events [21], geographical features [15], sounds [14] and categories with artificially created names [18]. These results show that basic level effects exist on a much wider scale than Rosch’s relatively clearcut examples, strengthening our claim that there is a need to automatically derive the basic level in large concept hierarchies.

The basic level is relatively universal since it is tightly coupled with universal physiological features such as what humans can perceive and what movements

² <http://cwi.nl/~hollink/basiclevelmtr2020/>

they are capable of [13]. That being said, it should be acknowledged that there are also individual factors that affect to what extent basic level effects occur. One of those factors is expertise. Domain experts may process subordinate levels with equal performance in the context of their domain of expertise [11,25]. Similarly, the familiarity of a person with an object plays a role, where familiarity increases basic level effects [24]. Finally, the prototypicality of an object is a factor; if someone perceives an object as a prototypical example of its class, basic level effects may increase [23]. These individual factors are outside the scope of the present paper, where we focus on the universal nature of basic level effects.

3 Related Work

The idea of a ‘basic level’ has been used in various applications that use conceptual hierarchies. In the context of the semantic web, it has been used in ontology creation [26,9], automatic ontology generation [7,4,3], ontology matching [8] and entity summarization [20].

Ordonez et al. [19] stress the importance of the basic level in computer vision. They propose a mapping between basic level concepts and the thesaurus of concept names that is used by existing visual recognition systems. Mathews et al. [16] use collaborative tags to predict basic level names of recognized objects in an automatic image captioning task.

For all these applications there is a need to identify which concepts are at the basic level. In the papers mentioned above this was done either manually [26,9], using heuristics [20,8], by looking at the frequency and order of occurrence of user generated tags [7,19,16], or using a measure called category utility [3,4].

The category utility [6] of a concept c is a measure of how well the knowledge that item i is a member of c increases the ability to predict features of i . For example, knowledge that i is a bird allows one to predict that i can fly, has wings, lays eggs, etc. Belohlavek and Trneck [1] compared the category utility measure for basic level prediction to two similar measures that were proposed earlier, such as cue validity [22] and Jones’ category-feature collocation measure [12], and found that they lead to similar predictions.

In contrast to category utility, cue validity and the category-feature collocation measure, our approach does not rely on the availability of explicit information about all features of a concept. In our experience, features such as “can fly” and “has wings” are seldom encoded in a concept hierarchy. Our approach builds on the idea of using tag frequency by including the frequency of occurrence of a concept in a natural language text corpus as a feature. Finally, our approach is inspired by some of the heuristics proposed before, e.g. with respect to the use of depth in the hierarchy [20] and lexical properties [8].

4 A Method for Basic Level Prediction

We cast the task of basic-level prediction as a binary classification problem: a concept in a hierarchy either is or is not at the basic level. In future work, we

intend to look into a multi-class classification task, distinguishing basic level, more specific and more generic concepts. Figure 1 shows an example hierarchy in which the basic level concepts have been identified. The figure illustrates that the basic level concepts can be at different levels in the hierarchy for different branches, and some branches may not include any basic level concepts.

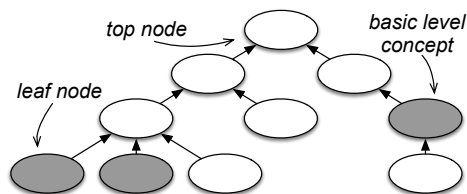


Fig. 1: Example hierarchy, basic level concepts in grey.

4.1 Extracting Three Types of Features

As input to the classifier, we extract structural, lexical and frequency features. We use the conceptual hierarchy to extract structural features about a concept. Rosch et al. [22] showed that at the basic level people tend to give longer descriptions when asked to describe a concept and were able to list most new attributes at this level. They concluded that the basic level is “the level at which categories carry the most information.” Based on this, we hypothesize that the amount of explicit information about a concept in a KOS can be used as a signal for basic level prediction. Accordingly, we collect the number of relations that the concept has in the KOS. Depending on the conceptual hierarchy, these can be is-a relations such as `rdfs:subClassof`, `skos:broader` or `wordnet:hyponym`. If other types of relations are present in the KOS, such as part-of relations, we count these as well. If the conceptual hierarchy contains natural language descriptions of the concepts, we include the length of the description as a feature. We also store the depth of the concept in the hierarchy, measured as the shortest path of is-a relations from the concept to the top node.

The use of lexical features is motivated by the fact that the basic level can be recognized in natural language. Many have claimed that basic level concepts are generally denoted by shorter and more polysemous words [18,25,8]. They are also the first words acquired by children [2]. The extraction of lexical features requires a mapping from concepts to words. In knowledge representation languages commonly used for KOSs, this mapping is for example given by `rdfs:label` or `skos:preferredLabel`. We extract the following lexical features for each concept: the length of the word(s) (measured in characters), the number of senses of the word(s) (i.e. polysemy) and the number of synonyms of the word.

Finally, we include the frequency of occurrence of a concept as a feature. Rosch’s initial experiments [22] demonstrated that people often choose basic level concepts when asked to describe an object, rather than subordinate or superordinate concepts. Therefore, we hypothesize that the frequency of occurrence of a word in a natural language corpus is a signal for basic level prediction.

Both lexical and frequency features are based on words. In many KOSs, one concept can be denoted by multiple words. Examples are the synonymous words *piano* and *pianoforte*, or *contrabass* and *double bass* (we treat multi-word phrases the same as single words). If this is the case, we need to aggregate multiple word-level feature values into one concept-level feature value. We use mean and minimum or maximum values for this purpose, depending on the feature.

4.2 Creating Manual Concept Labels for Training and Testing

We ask human annotators to manually label concepts as being basic level or not. An annotation protocol was provided that includes a short description of what the basic level is, as well as the results from Rosch’s initial experiments. For each concept, we provide the annotators with the synonyms that denote the concept, the position in the hierarchy, and a natural language description of the concept. The protocol lists additional sources that the annotator may consult: Wikipedia, Google web search and/or Google image search, all with a standardized query consisting of a word that denotes the concept and optionally a word of a superordinate concept. Finally, the protocol provides some hints that may help the annotator to decide on the correct label in case of doubt. For example, “at the basic level an item can often easily be identified even when seen from afar”, “at the basic level there is often a recognisable movement associated with the concept,” and “at the basic level images of the item often all look alike.”

5 Experiments and Evaluation on WordNet

We apply our basic level prediction method to WordNet, a lexical database of English [17]. It contains 155k words, organized into 117K synsets³. A synset can be seen as a set of words that denote the same concept. Synsets are connected to each other through the **hyponym** relation, which is an is-a relation. Each synset has a natural language description called a **gloss**. We train and test our approach on a subset of WordNet consisting of 518 manually labelled noun synsets.

WordNet presents a particularly interesting test ground considering (1) the depth of its hierarchy, making the identification of the basic level challenging and valuable, (2) its wide scope that includes all concepts that Rosch used in her original experiments, and (3) its widespread use: basic level predictions in WordNet are valuable for all knowledge graphs that have been linked to WordNet.

We extract lexical and frequency features from WordNet. For frequency features, we use Google Ngrams⁴, which provides data about how often a phrase – or “ngram” – appears in the Google Books corpus. This is expressed as the number of times a given ngram appears in a given year, divided by the total number of ngrams in that year. In the present study, the two most recent years of the Google Books corpus ‘English 2012’ were used, which comprises of 3.5K books in the English language. Table 1 details how each feature was operationalized. For feature selection details we refer to our preliminary work in [10].

³ <https://wordnet.princeton.edu/documentation/wnstats7wn>

⁴ <https://books.google.com/ngrams>

Table 1: Operationalization of structural, lexical and frequency features

Type	Feature name and operationalization
Struct.	nr_of_hyponyms Hyponym, the main is-a relation in WordNet, is transitive. We count the nr. of synsets in the complete hyponym-tree under the synset
Struct.	nr_of_direct_hypernyms Hypernym is the inverse relation of hyponym. As WordNet allows for multiple classification, some synsets have multiple hypernyms. We count the number of hypernyms directly above the synset.
Struct.	nr_of_partOfs The number of holonym plus meronym relations of the synset.
Struct.	depth_in_hierarchy The number of hyponyms in the shortest path from the synset to WordNet’s root noun <i>entity.n.01</i> .
Struct.	gloss_length The number of characters in the synset gloss.
Lex.	word_length_min The n.r of characters of the shortest word in the synset.
Lex.	polysemy_max The number of synsets in which the most polysemous word of the synset appears.
Lex.	nr_of_synonyms The number of words in the synset.
Freq.	G.Ngrams_score_mean the mean ngram score of the words in the synset.

5.1 Training and Test Set

The training and test set consists of synsets from three branches of WordNet: the complete hierarchies under the synsets *hand_tool.n.01*, *edible_fruit.n.01* and *musical_instrument.n.01*. In this paper, we will refer to these three hierarchies as “domains.” They correspond to three of the six non-biological hierarchies that Rosch reported in her seminal paper on basic level effects [22]. The WordNet hierarchies used in the present paper, however, are larger than Rosch’s experimental data; they consist of 150+ concepts per domain, whereas Rosch’s hierarchies consisted of 10 concepts each. What we call “domains” should not be confused with the topic-, usage- and geographical domain classification that WordNet provides for a small subset of synsets. All synsets in the training and test set were labelled by three annotators (the authors), with substantial inter-rater agreement (Krippendorff’s $\alpha = 0.72$). They labelled 518 synsets, of which 161 as basic level.

6 Results

6.1 Comparing Algorithms, Baselines and Annotators

We measure classification performance using a 10-fold cross-validation setup. In cases there the three annotators disagreed on the label, we use the majority vote. Table 2 lists (median) performance for five classifiers, which we implemented using an off-the-shelve toolkit⁵. We report Accuracy and Cohen’s Kappa (κ) – two measures commonly used for classification tasks – as well as Precision and Recall - two Information Retrieval measures, for which we consider basic level as the positive class. The best performing algorithm on all measures is the Random Forest, which we ran with the SMOTE algorithm to deal with class imbalance. Differences between algorithms are small and not in all cases significant.

⁵ The CARET Library in R <http://topepo.github.io/caret/index.html>

Table 2: Classification performance on entire training- and test set.

		Accuracy	Kappa	Precision	Recall
Classifiers: (median values)	LDA	0.81	0.59	0.73	0.74
	Decision tree	0.77	0.49	0.68	0.61
	K-nearest neighbors	0.70	0.37	0.59	0.63
	SVM	0.81	0.59	0.74	0.74
	Random Forest	0.82	0.61	0.75	0.76
Manual:	basic level at fixed depth	0.64	0.17	0.50	0.36
Randomly guessing:	all as basic level	0.36	0.00	0.36	1.00
	none as basic level	0.64	0.00	NaN	0.00
	50% as basic level	0.49	-0.02	0.35	0.49
	36% as basic level	0.54	0.01	0.37	0.37
Random Forest using labels: (median values)	of annotator 1	0.83	0.60	0.75	0.75
	of annotator 2	0.83	0.63	0.79	0.73
	of annotator 3	0.81	0.58	0.72	0.76
	on which all agreed	0.88	0.73	0.78	0.85
	majority vote	0.82	0.61	0.75	0.76

There are, to the best of our knowledge, no existing baseline algorithms that we can compare our results to. Instead, to place our results in context, we do two things. First, we compare the performance of the classifiers to an intuitive manual method that is based on the most important feature in the classification (as will be discussed in Section 6.2): depth in the hierarchy. We pick the level in the hierarchy with the highest number of basic level synsets (by looking at the training- and test set) and label all synsets at that level as basic level. This leads to a relatively high accuracy of .64 but a low κ of 0.17 (Table 2), both lower than any of the classifiers ($p \leq 0.01$). Second, we examine how far we would get with randomly assigning a percentage of the synsets to the basic level class: 100%, 0%, 50% or 36% (where the latter is the true percentage of basic level synsets in the data set). As expected, when we label all synsets as basic level, we achieve perfect recall; when we assign none of them to basic level, we achieve a high accuracy, which is on par with the accuracy of the manual method. All random guessing scenarios lead to a κ value of around zero (Table 2).

Next, we compare the three annotators, by looking at prediction performance when training and testing on manual labels given by annotator 1, 2 or 3. We find no significant differences here, which is good: it should not matter which annotator was chosen. Finally, we compare performance when training and testing is done on the majority vote of the annotators versus performance on only those synsets where all three annotators agree on the label (417 out of 518 synsets). We observe that performance is higher on the agreed labels, with median κ increasing from 0.61 to 0.73 (although this is not significant, $p = 0.06$).

To gain insights into what causes the observed difference between performance on agreed versus all synsets, we train a model on the agreed synsets (417 synsets), and test it on the synsets for which there was disagreement (101

synsets). When we evaluate this using the majority vote labels, κ drops to 0.06, an almost random classification. We hypothesize that concepts on which humans disagree are inherently difficult to classify because maybe there are no clear basic level effects in these cases. Concepts on which the annotators disagreed were, for example, rarely seen (by our annotators) fruits like the sweetsop, and the sibling concepts raisin, prune and dried apricot. The concept of berry was also a cause for disagreement, where one annotator labelled berry as basic level, while the other two labelled its hyponyms strawberry and blackberry as basic level. Future work will have to clarify whether basic level effects exist in these cases.

For brevity, in further experiments we only report κ , as this measure takes into account chance agreement [5], of the Random Forest and the manual method, trained and tested on majority vote labels. Other measures and the full results of the 10-fold cross validation will be part of the online supplementary data.

6.2 Basic Level Prediction Within and Across Domains

In Table 3a, we compare prediction performance of local models trained and tested on a single domain (Tools, Fruit, or Music) to performance of a global model on the entire data set (All). Results in Tools and Fruit are good (median $\kappa = 0.84$ and 0.79 resp.), while Music is more challenging (median $\kappa = 0.50$). The global model, with a median κ of 0.61, performs lower than most of the single domain models, suggesting that (some) features may not transfer well from one domain to another. This makes sense, as the distributions of feature values differ a lot over the three domains. For example, the mean gloss length is 20% longer in the Music hierarchy of WordNet. And, in our data set, part-of relations are rare except in the Fruit hierarchy. Table 3b lists feature importance in the global model and the single domain models, where the feature with the highest weight is ranked 1. The lists are relatively stable, with some marked differences, such as the importance of the gloss length and the number of partOf relations.

The manual method performs well on single domains (κ between 0.36 and 0.78, Table 3a) but badly when domains are pooled ($\kappa = 0.17$). Apparently, within a domain, basic level synsets reside largely at the same level; what this level is, varies per domain.

Finally, we examine what we consider the most realistic scenario: to predict basic level synsets in a new domain, for which we don't have manually labelled examples in the training set. To simulate this situation, we train on two domains, and test on a third. For example, we train on tools+fruit and test on music. Table 4 shows that performance of the Random Forest drops dramatically to κ values between -0.10 and 0.37 depending on which domain is considered as new. The manual method is even worse (κ between 0.02 and -0.42) because it relies on the level with the most basic level concepts, which is different in each domain.

To improve transfer learning, we include a per-domain normalization step: we divide the feature value by the mean feature value within the domain. Table 4 shows that normalization of structural features leads to a substantial performance gain (κ increases to 0.32-0.62 depending on the domain). Normalization of lexical or frequency features is not beneficial or even harmful to the results.

Table 3: Comparing local models (Tools, Fruit, Music) to a global model.

(a) Classification performance		(b) Features ranked in order of importance				
Random Forest (median (κ))		Feature	All	Tool	Fruit	Music
All	0.61	depth_in_hierarchy	1	1	1	2
Tools	0.84	G.Ngram_score	2	2	3	3
Fruit	0.79	gloss_length	3	4	4	1
Music	0.50	word_length_min	4	5	6	4
Manual method (κ)		polysemy_max	5	3	5	7
All	0.17	nr_of_partOfs	6	8	2	8
Tools	0.78	nr_of_hyponyms	7	6	8	5
Fruit	0.72	nr_of_synonyms	8	7	7	6
Music	0.36	nr_of_direct_hyperm.	9	9	9	9

Table 4: Performance (κ) in a new domain, with and without normalization.

		RF with normalization of features:				
Tested on: Trained on:		RF	Manual	Structural	Lexical	Frequency
Tools	Fruit+Music	0.37	0.02	0.62	0.43	0.34
Fruit	Tools+Music	-0.10	-0.42	0.41	0.06	-0.13
Music	Tools+Fruit	0.35	-0.01	0.32	0.21	0.34

6.3 Towards a Data Set of Basic Level Synsets in WordNet

We provide an RDF dataset of basic level concepts identified in the entire noun hierarchy of WordNet, using our approach with the best performing settings, i.e. a Random Forest trained on manually labelled synsets on which all annotators agreed, with per-domain normalization of structural features. With this data set, we aim to enable research into the use of basic level concepts in applications.

Per-domain normalization in a large knowledge graph like WordNet is non-trivial, as it requires a decision on what constitutes a domain. In our training and test set, this decision was intuitively easy: it consists of three disjoint branches, that correspond to three of Rosch’ high level categories. To split the 82K nouns in WordNet into domains is especially complicated due to many cases of multiple inheritance and its irregular structure with leaf nodes at every level of the hierarchy. We have implemented an ad hoc algorithm to split up WordNet into domains. It consists of 3 rules: (1) subbranches with between 50 and 300 nodes are treated as a single domain, (2) concepts with multiple parents from different domains are assigned to the smallest of those domains and (3) subbranches with less than 50 nodes are grouped together in a domain with their parent node.

To make the training set more representative, we manually labelled an additional 18 synsets; those synsets in the WordNet hierarchy that are between the three domains and WordNet’s top noun *entity.n.01*. Inter-rater agreement on this set was perfect, as all of them are by definition above the basic level.

The above results in an RDF data set of 10K basic level synsets.

7 Discussion, Future Work and Conclusion

We present a method to classify concepts from a conceptual hierarchy into basic level and not-basic level. We extract three types of concept features: lexical features, structural features and frequency features.

We show that, based on these features, a classifier can accurately predict the basic level within a single domain. The performance of a global model across multiple domains is slightly lower. Predictions in a new, previously unseen domain are meaningful only after a normalization step. Normalization can be straightforward, but does require a modularisation of the knowledge graph into domains.

A simple manual method – choosing a fixed depth in the hierarchy as basic level – gives reasonable results when applied to a single domain. When applied to a data set including multiple domains, or to a new domain, it doesn't produce meaningful results. We conclude that, also for the manual method, modularisation is a key aspect of basic level prediction. For small to medium size knowledge graphs, it may be feasible to do the modularization manually.

All three types of features proved important for the prediction. We believe that further improvements are possible from inclusion of additional frequency features. The Google Ngram scores that we used as frequency features gave a strong signal, and they did not need per-domain normalization. Other frequency features could include: the frequency of occurrence of words in specific corpora such as children's books or language-learning resources, and the frequency of occurrence of concepts in other conceptual hierarchies. WordNet contains also sense frequency counts, but they are available for less than 10% of our concepts.

A post-hoc discussion among the human annotators learned that for most concepts labelling was straightforward. A few cases were hard and annotators would have preferred to not make a choice between basic level or not. If it is true that there are concepts to which the basic level theory does not apply, it would be worthwhile to classify them as such, which would result in a classification into basic-level, not-basic-level and not-applicable. Concepts that are difficult to label for human annotators seem to be more challenging for the classifier as well. Applications will gain most from a correct classification of the straightforward cases, for which basic level effects can be expected to be strongest. In future work we intend to measure basic level effects for a larger set of concepts in a crowd-sourcing environment.

We ran our method with best performing settings on all noun concepts in WordNet and provide this as an RDF data set for reuse and further research.

References

1. Belohlavek, R., Trnecka, M.: Basic level in formal concept analysis: Interesting concepts and psychological ramifications. Proc. of IJCAI pp. 1233–1239 (2013)
2. Brown, R.: How shall a thing be called? Psychological review **65**(1), 14 (1958)
3. Cai, Y., Chen, W.H., Leung, H.F., Li, Q., Xie, H., Lau, R.Y., Min, H., Wang, F.L.: Context-aware ontologies generation with basic level concepts from collaborative tags. Neurocomputing **208**, 25–38 (2016)

4. Clerkin, P., Cunningham, P., Hayes, C.: Ontology discovery for the semantic web using hierarchical clustering. *Semantic Web Mining* **27** (2001)
5. Cohen, J.: A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* **20**(1), 37–46 (1960)
6. Corter, J.E., Gluck, M.A.: Explaining basic categories: Feature predictability and information. *Psychological Bulletin* **111**(2), 291 (1992)
7. Golder, S.A., Huberman, B.: The structure of collaborative tagging systems. *Journal of Information Science* **32** (2005)
8. Green, R.: Vocabulary Alignment via Basic Level Concepts. Final Report 2003 OCLC / ALISE Library and Information Science Research Grant Project (2006)
9. Hoekstra, R., Breuker, J., Di Bello, M., Boer, A.: The LKIF core ontology of basic legal concepts. In: *CEUR Workshop Proceedings*. vol. 321, pp. 43–63 (2007)
10. Hollink, L., Bilgin, A., van Ossenbruggen, J.: Is it a fruit, an apple or a granny smith? predicting the basic level in a concept hierarchy. arXiv:1910.12619 (2019)
11. Johnson, K.E., Mervis, C.B.: Effects of Varying Levels of Expertise on the Basic Level of Categorization. *J. of Experimental Psychology* **126**(3), 248–277 (1997)
12. Jones, G.V.: Identifying basic categories. *Psychol. Bulletin* **94**(3), 423–428 (1983)
13. Lakoff, G.: *Women, fire, and dangerous things*. University of Chicago press (2008)
14. Lemaitre, G., Heller, L.M.: Evidence for a basic level in a taxonomy of everyday action sounds. *Experimental Brain Research* **226**, 253–264 (2013)
15. Mark, D.M., Smith, B., Tversky, B.: Ontology and geographic objects: An empirical study of cognitive categorization. In: *International Conference on Spatial Information Theory*. pp. 283–298 (1999)
16. Mathews, A., Xie, L., He, X.: Choosing basic-level concept names using visual and language context. In: *2015 IEEE Winter Conference on Applications of Computer Vision*. pp. 595–602. IEEE (2015)
17. Miller, G.A.: Wordnet: a lexical database for english. *Communications of the ACM* **38**(11), 39–41 (1995)
18. Murphy, G.L., Smith, E.E.: Basic-level superiority in picture categorization. *Journal of Verbal Learning and Verbal Behavior* **21**(1), 1–20 (1982)
19. Ordonez, V., Deng, J., Choi, Y., Berg, A.C., Berg, T.L.: From large scale image categorization to entry-level categories. In: *Proc. of ICCV*. pp. 2768–2775 (2013)
20. Peroni, S., Motta, E., d’Aquin, M.: Identifying key concepts in an ontology, through the integration of cognitive principles with statistical and topological measures. In: *Proc. of the 3rd Asian Semantic Web Conference, ASWC*. pp. 242–256 (2008)
21. Rifkin, A.: Evidence for a basic level in event taxonomies. *Memory & Cognition* **13**(6), 538–556 (1985). <https://doi.org/10.3758/BF03198325>
22. Rosch, E., Mervis, C.B., Gray, W.D., Johnson, D.M., Boyes-Braem, P.: Basic objects in natural categories. *Cognitive Psychology* **8**(3), 382–439 (1976)
23. Rosch, E., Simpson, C., Miller, R.S.: Structural bases of typicality effects. *J. of Experimental Psychology: Human Perception and Performance* **2**(4), 491–502 (1976)
24. Smith, E.: Effects of Familiarity on Stimulus Recognition and Categorization. *Journal of Experimental Psychology* **74**(3), 324–332 (1967)
25. Tanaka, J.W., Taylor, M.: Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive Psychology* **23**(3), 457–482 (1991)
26. Uschold, M., King, M.: Towards a Methodology for Building Ontologies. *Methodology* **80**(July), 275–280 (1995). <https://doi.org/10.1.1.55.5357>
27. Zeng, M.: Knowledge organization systems (kos). *Knowledge Organization* **35**, 160–182 (01 2008). <https://doi.org/10.5771/0943-7444-2008-2-3-160>